

AUTOMATIC PRODUCT REVIEW SENTIMENT ANALYSIS USING VADER AND FEATURE VISUALIZATION

HARISH RAO M¹ & SHASHIKUMAR D.R.²

¹Research Scholar, M. Tech (CSE), Cambridge Institute of Technology, Bangalore, Karnataka, India

²Head and Professor, Department of Computer Science, Cambridge Institute of Technology, Bangalore, Karnataka, India

ABSTRACT

Sentiment polarity analysis has been a popular research field for data scientists over the last decade. Movie reviews, hotel reviews, social media like twitter reviews and product reviews have been the subjects of sentiment polarity analysis. NLTK has been facilitating these researchers with necessary classification tools to verify and fine-tune the accuracy of sentiment polarity analysis models. The most interesting part of the research is the sentiment polarity using the intensity of the sentiments in the reviews. The Vader sentiment analysis tool is one such tool which uses a specially developed lexicon to classify the sentiment based on the intensity of sentiments. Vader also facilitates unsupervised sentiment analysis, unlike other supervised machine learning techniques. This study explores Vader tool for unsupervised and online sentiment analysis of product reviews. The study also focuses on the domain-based training datasets and their universal applicability for sentiment classification. Finally, the study highlights the usefulness of direct visualization techniques for selected high frequency negative and positive feature sentiments.

KEYWORDS: Sentiment Polarity, NLTK Tools, Unsupervised Learning, Vader, Training and Testing, Datasets, Visualization Techniques & Feature Sentiment Polarity

Received: Jun 24, 2017; **Accepted:** Jul 15, 2017; **Published:** Jul 28, 2017; **Paper Id.:** IJCEITRAUG20178

INTRODUCTION

Direct or online sellers of products and services facilitate their customers to express their sentiments and rating online for their products and services. Besides them, there are independent websites like mouthshuts.com, rotten tomatoes, imdb, epinion.com, tripadvisor.com, trivago.com. And many others, which consist of customer reviews and star ratings. Social media like twitter and Facebook also consist of users expressing their sentiments on a variety of subjects, of which include politics, religion, many general and social issues, and also product and service related issues. Bo Pang and Lee [1] contributed to the sentiment polarity analysis of different domain and generated large datasets available online for other researchers pursuing the sentiment polarity analysis in different domains. There are several challenges posed to the researchers in sentiment analysis. Sentences and words used in the sentiments expressed can be tricky and classification techniques may have too many false positives and false negatives in sentiment classification. Sentiments expressed by a single customer may extend to several paragraphs. A single sentiment may have both positive and negative features as well as just a neutral opinion suggesting some expected improvements in the product and services. Sentiments can also indicate the intensity of sentiment by certain select type of words and exclamatory symbols as well as visual symbols like Emoticons. Hence, researchers turned their attention to the intensity of the sentiments to improve the classification of sentiment polarity. Lexicons have been developed to capture the sentiments for intensity. Bo Pang and Lee have produced datasets with

annotation and labelling of the sentiments with plus and minus numerical rating scales. Unsupervised machine learning techniques have the disadvantage of long computational time taken for analysis and manual annotation of sentiment polarity is tedious. The newly developed Vader sentiment analysis tool [4] facilitates unsupervised approach as well as the lexicon developed to capture the sentiment intensity based classification using probability of compound nature of the sentiments. The tool provides for adjustment of the parameters to improve the accuracy of sentiment intensity based classification of polarity of sentiments. The focus of this work is to study the effectiveness of the Vader tool for unsupervised online sentiment polarity analysis. Since it is not enough to say that overall sentiment on a product or service is negative or positive, researchers turned their attention to significant features and aspect extraction based on the frequency of negative or positive sentiments expressed on them. Techniques again required various language processing tools to clean the collected reviews for irrelevant words and expressions and application of tools such as a bag of words and n-gram association rules to capture word level sentiments. The researchers on sentiment polarity or sentiment orientation use more than 20 different methods and as well as many different classifiers to get more accurate predictions. Prediction accuracies are calculated through precision, recall and f-score metrics in all cases.

SUPERVISED MACHINE LEARNING

Multi-domain dataset availability online helped many researchers to put their analytical minds to the sentiment analysis and classification, Movie reviews, Hotel reviews, twitter reviews as well as Amazon product reviews are major sources for trying out different tools and techniques to improve prediction accuracies and lowering computation times. New algorithms and methods are constantly being searched out to give more reliable sentiment classifications for different domains. There are many survey papers and project reports which enlist the significant amount of work on sentiment analysis. We review briefly those publications which have significance and relevance to this paper.

Lopa mudra et.al [5] applied unsupervised Naïve Bayes(NB) and K-Nearest Neighbors (K-NN) algorithms to movie and hotel reviews. Accuracy of movie reviews from each of these techniques found was 82.4% (NB) and K-NN (69.8%). But hotel reviews gave lower accuracy of 55.1% (NB) and 52.1%(K-NN). They found that accuracy improves in proportion to the data size. They found that NB is more accurate, even for smaller data sets as compared to K-NN. MD. Shad Akthar et.al [6] used a hybrid approach using NLP, Linear regression (LR) and SVM tools to analyze twitter sentiments, restaurant and laptop reviews. They applied convolutional neural network algorithm coupled with SVM and found that accuracy for twitter and movie reviews ranged from 44.9 to 62.5%. For restaurants and laptops accuracies improved to 77.16% and 68%. Apple et.al [7] used a hybrid approach, including semantic rules, fuzzy sets and enriched the lexicon. They have compared this approach with NB and Maximum Entropy methods and found that their method gave higher accuracy (0.76) as compared to NB (0.67) and Maxent (0.76) when applied to movie review datasets. Abbasi et.al [9] attempted to do benchmarking of twitter sentiment analysis tools. 20 tools were applied to 5 test beds like telecom, pharma, retail, security and they have used 3-class polarity of data into positive, negative and neutral sentiments. Out of the 20 tools Sentistrength reported the highest accuracy of 67.5% and Pharma and Telco test beds reported highest accuracies of 74.7% and 71.3% for Sentistrength tool. They have used 5 workbench tools whose average accuracies ranged from 66.9 to 71.4%. Using ataxonomy based root cause analysis they found maximum errors arose from sarcasm, modifiers, jokes, rhetoric as well as irrelevant positive and negative sentiment categories. They recommended error analysis, annotated twitter data sets for better accuracy performance. Abhijit et.al [39] performed sentiment analysis on two movie review datasets and one dataset detecting insults in the user's comments. Effect of tokenization by porter stemmer and POS

tagging on accuracy was studied. Porter stemmer tagging has given better accuracies of 0.87(NB) and 0.89(SVM). Vivek et.al [11] used an enhanced NB model applied to movie datasets. NB according to them provides several orders of magnitude lower in time as compared to vector models. They have used NB Bernoulli algorithm for classification. Laplacian smoothing applied over NB to handle the case when classifier encounters a word that does not exist in the training set. Another measure they took to improve the accuracy of their model is through appropriate negation handling. They have also applied n-gram classification techniques in the analysis. The reported accuracies for NBB and bi-gram are 82.8% and 85.2%. They found no significant improvement above bigram association of features. Kafir Bar [23] did sentiment analysis on movie and twitter datasets using NB, SVM (poly & Linear) and K-NN. NB and SVM gave same accuracies of (78 to 84%) for both datasets while K-NN gave poor accuracies (0.62 to 0.64). Bo Pang et.al [1] the leaders in sentiment analysis in one of the excellent research papers made an important observation by experimenting human classified sentiment polarity that humans also do not exceed an accuracy of 0.69. This is considered as a benchmark for assessing machine learning accuracies for performance comparison. They have used movie review corpora and NB, ME and SVM techniques for sentiment analysis. For feature analysis, they have applied unigram and bigram association techniques. Features are analyzed by counting the frequency and presence. ME did not give any better accuracies as compared to NB and SVM. SVM and NB produced almost same accuracies. Frequency based classification of features produced slightly lower accuracy as compared to presence based feature classification. Prateek in his master's thesis [15] conducted unsupervised studies on twitter sentiment on 3 major political parties in India. He used movie corpora as training datasets. He applied 6 different classifying techniques for training with movie datasets to classify the twitter sentiment data. Multinomial and Bernoulli NB resulted in better accuracies as compared to other 4 methods. LR and LSVM reported lower accuracy levels around 71-72%. Xinmiao et.al [19] tried global optimization approach to multipolarity sentiment analysis. They found that 3-class polarity classification resulted in much lower accuracies as compared to 2-class polarity classification.

UNSUPERVISED SENTIMENT ANALYSIS-VADER SENTIMENT INTENSITY ANALYZER

This paper relates to unsupervised product review, analysis and we therefore review here the significant contributions found in the literature for supervised as well as unsupervised sentiment polarity analysis on product reviews. Among the many such contributions one best research work found in the excellent paper by Ribeiro et.al [8] They have done a comprehensive benchmark comparison of state-of-the practice sentiment analysis methods. 24 popular methods are included in the benchmark study. Vader is one among the chosen methods. They used 18 benchmark datasets covering most popular domains. They have included both 3-class and 2-class sentiment polarity classification of datasets. They found that in the social media context Vader stood in the third place in ranking for 3-class experiments and it took fifth to ninth rank when applied to other domain datasets and 2-class polarity experiments. Vader, however performed the best among the unsupervised methods which included SO-CALL and USENT. On tweet datasets Vader had accuracies of 84.4 and 99%. The study highlights that there is no single consistent accuracy, precision and ranking of any of the 24 methods applied in the study. The authors recommended research to improve these methods to raise their accuracy levels and consistency. Narendra and Samik [21] used non-parser dependent and semantic role labelling. They used a binary classification method of aspect tagging to differentiate a word as aspect or a non-aspect. Their analysis of two products from Amazon reviews dataset to yield precision, recall, f1score and accuracy. While Accuracy is impressively high around 97.2 % and 97.4 % for Zen 40 GB and Apex DVD, f1 scores were very poor around 38.4 % and 36.7%. Qian et.al [34] published data on their experiments with semantic similarity and aspect association algorithms on amazon electronic data sets. They found good F1

scores for algorithm using aspect association. With additional knowledge extracted from unsupervised review datasets based recommendations. Nikhil et.al [22] used WEKA-3 classifier on hand annotated amazon review datasets for attribute extraction. They inspected the hand annotated datasets and found inconsistency, incomplete and wrong annotation in them. They found poor precision around 13 to 16% while recall scores were ranged from 78 to 89% for original annotated features in the datasets. When they tried to prune the misannotated features from original datasets they found no improvement in precision while recall scores dropped to 34 to 57%. The most important result from this work is that when training and test datasets are from the same product all metrics are impressively high. In a 90/10 split train/test data on same product CANON, accuracies recorded were 96.65% and other metrics were impressively good. When training set is Canon and test set is Nikon, accuracy scores are good around 88.73% and other metrics like precision, recall and F1 score were found to be lower ranging from 54% to 61%. From these results they concluded that feature selection based on frequency of occurrence and POS tagging are important to get the best classification metrics. Subhabratha et.al [24] applied rule based feature extraction classification on domain specific training amazon data sets and found varied accuracies for 14 different products ranging from 57.6% to 78.6%. If they eliminated domain specific implicit or hidden features for classification accuracies improved to 83% and 87% for two select products like Camera and mobile phone. Maria et.al [25] used textblob in python library for sentiment classification for amazon product review database using a polarity scale for 3-class classification. Multinomial NB and SVM used as classifiers. Train and test split is performed with 50/50 split on data. MNB gave accuracy of 72.95% but took only 0.13 seconds for analysis while SVM gave better accuracy of 80.1% with longer computational time of 16 minutes and 38 seconds. One significant fact on this work is that they used word clouds to visualize the review features for the products. Turney et.al [14] analyzed unsupervised semantic orientation on opinion collected review data sets to four different domains consisting of automobiles, movies, travel destinations and banks. They used 400 reviews where opinion users gave 41% negative recommendations 59% positive recommendations. Automobiles and banks have good accuracies (84% and 81% as compared to movies scored only 65 to 66.7%. Travel destination gave 64% and 80.6% accuracy for two different destinations. Richa et.al [28] used a dictionary based unsupervised technique to classify sentiment orientation on amazon reviews collected from Amazon site for mobile phones. They collected metrics for different features like design, battery, camera, processor, cost, ease of use and miscellaneous. The significance of this study is that it a human based classification on all the collected reviews is used to compare with the unsupervised machine learning technique used in the study. The features like design, processor, cost and ease of use scored 0.68 to 0.81 in all the metrics. Camera and battery features scored only 0.5 in accuracy. Akshay and Navjyothi [29] studied the impact of non-domain specific and domain-specific ontology for sentiment classification. They found small improvement in accuracy of 62% to 65% for computers by using domain specific ontology. News articles scored accuracy of 81% and 84% and shopping comments scored 59.4 and 62.5 in accuracy for non-domain specific as well as domain specific ontology. But in each case 3% improvement was found with use of domain specific ontology. Hang et.al [27] tried an unsupervised passive aggression algorithm and compared the performance with other traditional algorithm and produced good average precision, recall and accuracy scores but negative class scores are low around 0.6 to 0.7 with their new algorithm also. Therese et.al [30] adopted a method to crawl the product urls directly and put to semantic analysis the data retrieved by offline method. They recommend this approach so that users can have direct information on semantic orientation while crawling for product reviews from e-commerce site. But their work used online data collection and offline semantic analysis. Gurneet and Abhinash [31] extracted top 10 reviews from flipkart site for MotoX phone and subjected them to sentimental analysis using POS tagger and NB. They got impressive accuracies well above 90% for precision, recall and f1 score. A novel rule based approach that depends on common sense knowledge and sentence dependence trees applied by Soujanya et.al [32] to

detect the explicit and implicit aspects. They used Hu and Liu provided dataset for explicit aspect extraction algorithm and Semeval2014 dataset and for implicit aspect extraction the dataset used was the corpus developed by Cruz-Garcia 2014. They applied their technique to DVD player, Canon G3, Jukebox, Nikon Coolpix, and Nokia 6610 datasets provided by Hue and Liu. They reported only precision and recall which were impressively higher above 90% for all the investigated products except for Nikon Coolpix having the scores of 82 % and 86% respectively. They applied this approach to the products like laptops and restaurants from Semeval2014 datasets with precision and recall scores between 82 to 88% with restaurant dataset giving better scores than laptops. Particle swarm optimization algorithm named as PSOGO-senti was developed and applied by Xinmiao et.al [19] specially applied to Chinese sentiment analysis for binary and multi polarity sentiment classification. They found that this algorithm is capable of eliminating redundant and noisy features. They got precision, recall and f1scores of 0.9, 0.69 and 0.52 for 2-pol, 3-pol and 5-pol experiment in that order when applied to Ctrip and Guahao datasets. Arbolede et.al [33] used visualization techniques like word trees using Stanford parser API which are more informative about the feature sentiments than classifier based extracted features. Sherin and Shine [40] used a modified AAACV technique applied to amazon review datasets on cameras. This is an unsupervised method without any training data sets used. They got accuracy scores of 68, 74.3 and 73.1% for 3 different camera datasets with 300 reviews. In a master's thesis by Yanyan [35] used two different domain datasets (Movies and Amazon product reviews) and 3 different lexicons (subjective clues, wordnet, opinion words for product reviews). Both unsupervised and supervised techniques studied by him. Intensification rule method scored in the range of 0.53 to 0.7 for precision, recall, and f1 score for 5 different camera products from Amazon review datasets. While sentence relation based feature, extraction yielded 0.57 to 0.76 for the same products and measures. In a fine grain analysis done by this author revealed that neutral and negative product feature classification were affecting overall scores. Elliot and Zachary [36] presented their study on laptop and restaurant datasets and found that by correcting aspect-sentiment pairs they could get accuracy of 78% for both datasets. Feature based summarization (FBS) technique is used by Minqing and Bing [37] applied to 4 different product datasets from Amazon review (Camera, Mobile, MP3, and DVD). They published Recall and precision data for association mining, compactness pruning, P-support pruning and infrequent feature identification. The scores ranged from 0.56 to 0.80. Xing and Justin [38] applied sentence level as well as review level categorization on amazon online product review datasets. At sentence level f1 scores were 0.80 and at review level F1score were 0.73. Their results also show that NB performed equal or better than SVM and Random Forest (RF) methods.

ISSUES TO BE ADDRESSED IN SENTIMENT ANALYSIS RESEARCH

From the large amount of data, one can deduce that sentiment classification and feature and aspect analysis is in a constant state of flux with no single corpora, lexicon, algorithm, classifier, tagger, levels of sentiment analysis gave a consistent accuracy as well as all other metrics. Vader is one of the top 10 ranked sentiment classification tools and is found to give better performance than other unsupervised methods like SO-CALL and USENT. Vader is powerful lexicon yet to be exploited. It is faster and accurate enough to give sentiment orientation and feature classification.

Most existing work either used available corpora as training and testing datasets or collected data and analyzed offline with unsupervised machine learning or unsupervised classification methods. It will be more useful to consumers as well as an eCommerce retailer if reviews could be classified live online rather off-line. Equally feature extraction and aspect extraction results vary with respect to noisy and redundant data. Whether to use frequency and or presence is still not been given consistent ideas from the available literature. Very few have tried powerful visualization methods which can give the

features with their connected sentiment indirect word tree form.

Unlike movie, hotels and twitter sentiment one can see that product reviews and especially related to electronics products are highly positively biased. This can be seen from published Amazon product reviews as well as published data in the product review sentiment analysis related work in the past. This is expected because the customer goes ahead to buy these products after getting word of mouth assessment as well as published reviews and star ratings on eCommerce sites. Hence, they already have positive orientation the product, but they will give negative sentiments if any feature did not meet their expectation or any unexpected failures in the product or serviceability. Hence product review sentiment analysis is not going to add to already known overall sentiment on the product, but feature related sentiment extraction and visualization will be more helpful to customers as well as producers as well as sellers.

In this work, we have been successful to use the Vader online product review and unsupervised sentiment classification. Another point is how important are domain specific training datasets for achieving the required accuracy of classification. We have tried to study this aspect also. More importantly, this study shows how data visualization tools can give more information on a feature or aspect based sentiments to get valuable clues to customer sentiment related to specific features

EXPERIMENTAL FRAMEWORK, TOOLS AND TECHNIQUES

Web crawling, Product review data extraction is done using open source Python 2.7 and its package rich library tools. For Natural language processing NLTK tools downloaded and used. Vader tool integrated with NLTK with its Sentiment Intensity Analyzer (SIA) has been used for unsupervised sentiment classification coupled with automatic crawling and product review extraction. From Python library, SCIKIT Sclera based Naïve Bayes classifier and Linear modelling modules imported and applied for training and classification and SKLEARN metrics are employed for measuring the precision(P), recall(R), F1score(F1).

Two different Corpora data sets used in this study are well used by most researchers and cited by them in their papers.

- Amazon review data sets by M. Hu and B. Liu (2014). DVD, MP3, Digi camera, Mobile phone data which is pre-annotated by the authors.
- Amazon review data published by Julian McAuley, UCSD (<http://jmcauley.ucsd.edu/data/amazon/>).
- Labelled and classified movie dataset published by Bo. Pang et.al (<http://www.cs.cornell.edu/people/pabo/movie-review-data/>).

The movie data set is used for training sets for feature sentiment polarity extraction. For testing domain specific feature sentiment classification, Julian data sets used as domain specific training for testing data sets for Laptops, Mobiles and Cameras.

Vader lexicon is down loaded to NLTK library. Vader assigns a compound probability for each review which is an important parameter to be modified to suit to the requirement of analyzer by trial and error using human evaluation of for manual classification and comparing with Vader-SIA. Default values of this are positive (>0.05), negative (<-0.05) and Neutral (Between -0.05 and 0.05). For our study, we replaced these by manual reviews of the classified posts for most

accurate classification and replaced 0.05 by 0.2 for same 3-level classification by Vader SIS and got the best results for Product reviews collected from Amazon and Flipkart sites.

A web crawler is coded in python will crawl the product review pages and retrieve up to 10 to 15 pages of reviews due to the limitation set by the web site administrators. As each page is crawled each review is retrieved cleaned and stored as a post in a text file as well as automatically taken into Vader SIA to classify each post online as positive, negative or neutral Following Vader classification all the positive and negative review data sets are processed through two different (NBB and LR) classifiers simultaneously for feature based sentiment analysis and classification as depicted in figure2. Following Vader classification all the positive and negative review data sets are processed through two different (NBB and LR) classifiers simultaneously for feature based sentiment analysis and classification as depicted in figure2.

```

Terminal
+ {'neg': 0.0, 'neu': 0.256, 'pos': 0.744, 'compound': 0.4404}
X {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.4215}
{'neg': 0.0, 'neu': 0.417, 'pos': 0.583, 'compound': 0.4215}
{'neg': 0.116, 'neu': 0.838, 'pos': 0.046, 'compound': -0.8035}
{'neg': 0.0, 'neu': 0.408, 'pos': 0.592, 'compound': 0.7003}
{'neg': 0.213, 'neu': 0.787, 'pos': 0.0, 'compound': -0.6597}
sentiments computed for 190 posts
('total posts =', 190, 'positives =', 116, 'negatives =', 43, 'neutrals =', 31)

```

Figure1: Typical Console Output after Online Classification

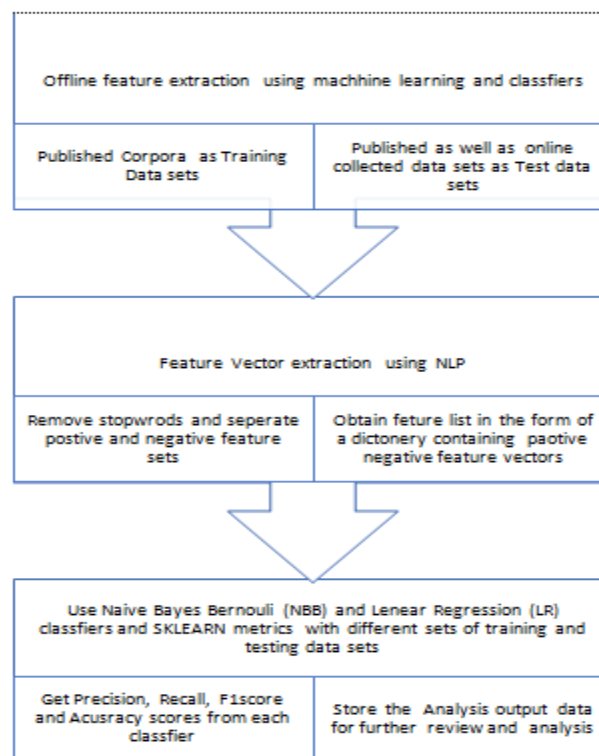


Figure 2: Feature Selection and Classification

RESULTS AND ANALYSIS

Cross-Domain Training on IMDB Movie Datasets (Pang and Lee)

In the first set of experiments we tried to use published polarity data for Movie datasets (cross-domain) and for product review datasets, as test data we have tried Hu& Liu datasets, Julian datasets and our collection. The corresponding classification metrics are presented in tables 1 to 3

Table 1: Trained With Movie Data Set and Tested on Hu & Liu Product Review Datasets

Training Cross Domain Products (Hu&Liu)	Naïve Bayes Classifier				Linear Regression Classifier			
	Precision	Recall	F1score	Accuracy	Precision	Recall	F1score	Accuracy
APEX DVD player	0.964	0.66	0.783	0.667	0.954	0.671	0.788	0.680
CANNONG3	0.979	0.811	0.887	0.803	0.959	0.814	0.881	0.795
CREATIVE-ZEN40B	0.963	0.729	0.830	0.726	0.927	0.735	0.819	0.717
NIKONcoolpix4300	0.956	0.802	0.873	0.779	0.934	0.799	0.861	0.762

Table 2: Trained With Movie Data Set and Tested on Julian Product Review Datasets

Training Cross Domain Products (Julian)	Naïve Bayes Classifier				Linear Regression Classifier			
	Precision	Recall	F1score	Accuracy	Precision	Recall	F1score	Accuracy
Laptops	0.805	0.912	0.855	0.761	1.000	0.790	0.882	0.790
Mobiles	0.818	0.893	0.854	0.759	0.934	0.799	0.861	0.762
Cameras	0.795	0.963	0.871	0.778	0.948	0.809	0.873	0.784

Table 3: Trained With Movie Data Set and Tested on Our Collection of Product Review Datasets

Training Cross Domain Products (Our Collection)	Naïve Bayes Classifier				Linear Regression Classifier			
	Precision	Recall	F1score	Accuracy	Precision	Recall	F1score	Accuracy
ApplemacbkAir	0.986	0.921	0.952	0.911	0.958	0.925	0.941	0.892
Lenovo-Ideapad110	0.899	0.778	0.834	0.750	0.860	0.799	0.828	0.750
CANON-EOS700	0.985	0.918	0.950	0.905	0.977	0.920	0.948	0.902
SONY-DSCH300	0.946	0.926	0.936	0.884	0.946	0.932	0.930	0.891
SAMSUNG S7	0.968	0.872	0.917	0.855	0.962	0.871	0.914	0.850
SAMSUNG J7	0.968	0.804	0.866	0.779	0.939	0.804	0.866	0.779

Domain Specific Training on Julian Datasets Like Laptops, Mobiles and Cameras

In the second set of experiment we have tried the training set from *domain specific electronic* product dataset collection of Julian (laptops, cameras and mobiles) and corresponding domain specific products as training data sets from Hu&Liu collection as well as our own collections. The results are presented in tables 4 and 5

Table4: Trained With Julian Product Review Data Set and Tested on Hu&Liu Product Review Datasets

Training Domain Specific Products (Hu&Liu)	Naïve Bayes Classifier				Linear Regression Classifier			
	Precision	Recall	F1score	Accuracy	Precision	Recall	F1score	Accuracy
CANONG3	1	0.79	0.882	0.789	1.000	0.790	0.882	0.790
NIKON4300	0.956	0.803	0.873	0.779	0.934	0.799	0.861	0.762
NOKIA6610	0.974	0.803	0.880	0.793	0.948	0.809	0.873	0.784

Table 5: Trained With Julian Product Review Data Set and Tested on Our Collection of Product Review Datasets

Training Domain Specific Products (Hu&Liu)	Naïve Bayes Classifier				Linear Regression Classifier			
	Precision	Recall	F1score	Accuracy	Precision	Recall	F1score	Accuracy
ApplemacbkAir	0.951	1	0.907	0.858	0.951	1	0.907	0.858
Lenovo-Ideapad110	0.923	1.000	0.856	0.647	0.923	1.000	0.856	0.647
CANON-EOS700	1.000	0.912	0.954	0.912	1.000	0.912	0.954	0.912
SONY-DSCH300	1.000	0.891	0.942	0.891	1.000	0.891	0.942	0.891
SAMSUNG S7	0.909	0.967	0.833	0.760	0.909	0.967	0.833	0.760
SAMSUNG J7	0.992	0.714	0.998	0.860	0.992	0.714	0.998	0.860

For such as complex task as sentiment polarity analysis, where multiple algorithms, lexicons, text processors, and classifiers are used to predict sentiment polarity by both supervised and unsupervised methods, an accuracy level of 0.8 to 0.90 considered gold standard as it exceeds human manual classification accuracy from that context VADER sentiment polarity classifier performed very well with the most of the tested product reviews with an accuracy between 0.8 to 0.9.

When we tried to study the impact of domain specific training sets against the cross-domain specific training datasets, we do not find any great difference in accuracy metrics but positive precision scores are impressively high reaching 1.0 even. This will be expected because product reviews are more positively biased and are sensitive to domain specific sentiment polarity.

Feature Sentiment Visualization

There are good tools to convert documented text to word clouds and word trees. These visualizing tools are not frequently applied to this type of analysis. We would like to show a sample cases of these visualization in order not to exceed the prescribed length of this paper. We have applied python library package for word cloud building and an open source java application called jigsaw for word tree visualization of featured sentiments

First, we feed the positive or negative review texts for each product to word cloud application to form a word cloud as shown in figure

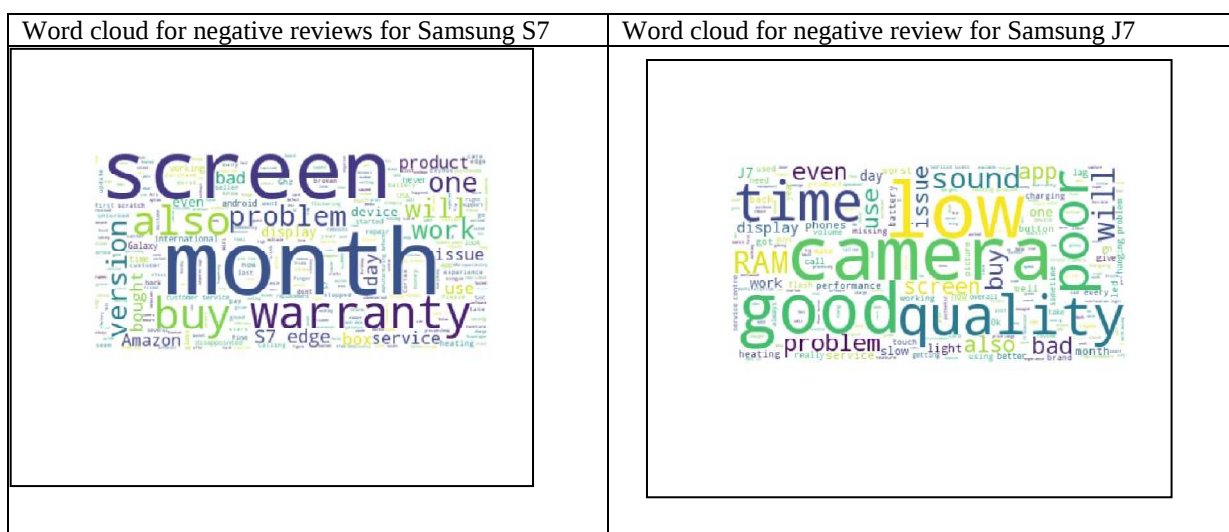


Figure 3: Negative Word Clouds for S7 Edge and J7 Samsung Mobile Phones

Intensity and size of the feature or aspect word is proportional to the frequency of its appearance in the reviews

For example, in case of S7 **screen** is highest frequency sentiment where as J7 **camera** is the highest frequency sentiment. We also should take note of adjectives like good, quality, poor, bad which can reveal some features annoying or delighting the customer. Now we will see word trees for the features like screen, good shown in the word cloud.

Now we will feed the same text file to a tool called JIGSAW which can provide us the word tree surrounding that keyword or feature with frequency of its occurrence in the reviews. The two corresponding word trees for screen (S7edge) and camera (J7) are shown in the figure

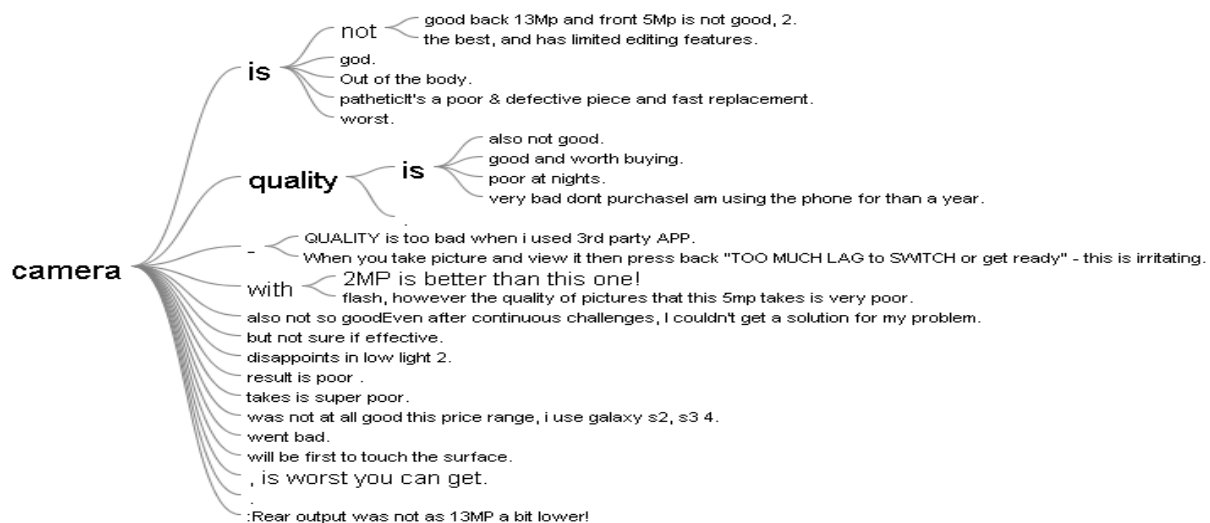


Figure 4: Jigsaw Word Tree for Keyword Camera for Samsung J7 Mobile Phone

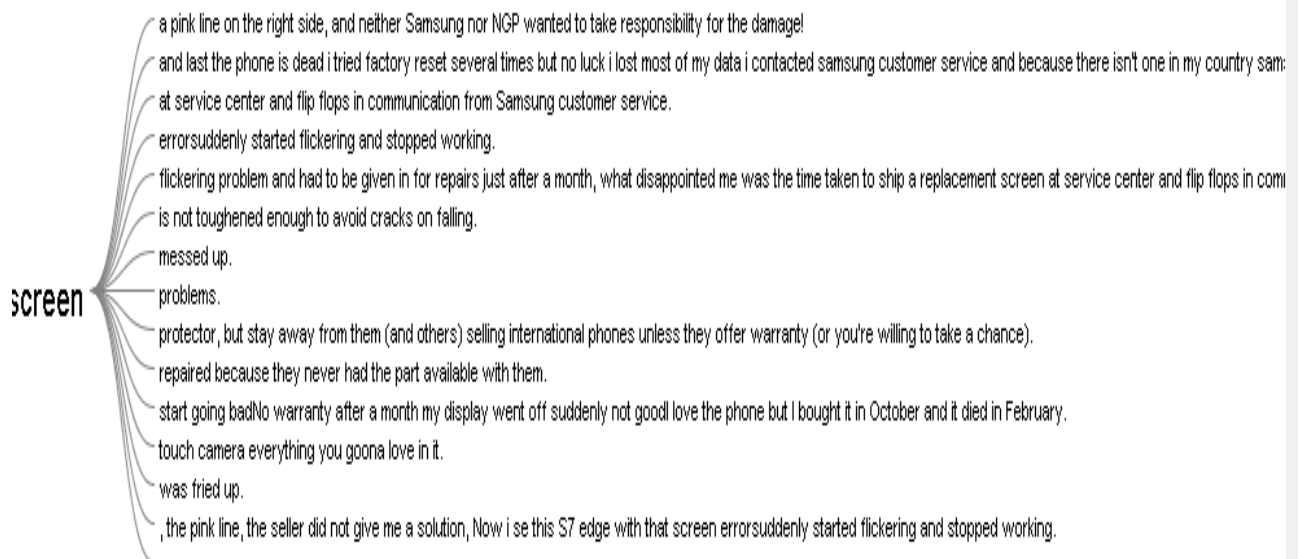


Figure 5: Jigsaw Word Tree for Keyword Screen for Samsung S7 Edge Mobile Phone

It is interesting to see a positive word appearing in S7 edge negative cloud and when we projected this word

“good” into a word tree through jigsaw all sentiments surrounding this word are negative and sarcastic. Thus, VADER proves to be good to pick up contextual polarity despite the positive word (good) association. This can be seen in figure

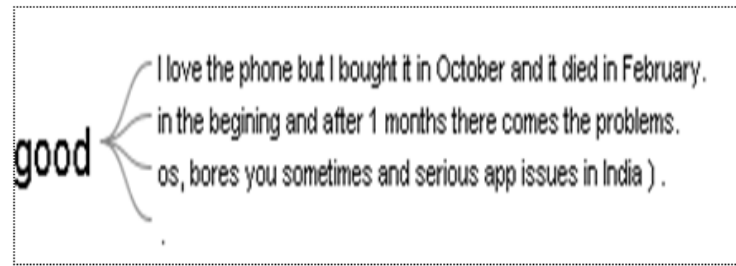


Figure 6: Word Tree for Key Word Bad for S7

The feature sentiments extracted in word trees from negative reviews of 5 products presented in table-6

Table 6: Extracted Feature Sentiments from the Word Trees for the Negative Reviews of 6 Products

Product	Keyword in Negative Word Cloud	Aspect Highlighted in Word Tree
Lenovo-ThinkPad	Battery	Charging and discharging times are disappointing for customers
Samsung S7 edge	Screen,	The screen has pink lines often and display is dead in a new phone
Samsung J7	Camera	The front 5mp camera is having bad quality pictures
Canon EOS 700	Good	The product is good, but Battery is not good
Sony DSCH 300	Camera	The camera has multiple issues

CONCLUSIONS AND FUTURE WORK

For product reviews, it is more important to have feature level sentiment polarity analysis. Online reviews with unsupervised sentiment classification require fast and effective sentiment analyzer. We found Vader a powerful tool not fully explored and exploited for product review sentiment analysis while it has been used mostly for social media sentiment analysis. Vader tested with both IMDB movie review training sets as well as domain specific Julian training set same levels of good accuracy. Vader scored accuracy levels closer to most benchmark studies (0.80 to 0.90).

Future work should explore to develop a web based application with appropriate authorization from eCommerce sites for automatic crawling with integration of Vader analyzer as well as visualization tools. Customers, producers an online retailer get valuable information about product feature and aspect level sentiment orientation about a product of interest in a fast and accurate manner

REFERENCES

1. Bo.Pang, and Lillian.Lee, “Proceedings of the Conference on Empirical Methods in Natural language processing (EMNLP)”, Philadelphia, July, 2002, pp.79-86
2. Bing.Liu, “Setmtimentanlysis and opinion mining ” Morgan & Claypool Publishers.
3. Nipun.M,Sashikant,K. and Priyank.P, stanford university project report,” Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews”web.stanford.edu/class/cs276a/projects/reports/nmehra-kshashi.
4. C.J.Hutto,and E.Gilbert, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” Association for the Advancement of Artificial Intelli-gence 2014.

5. Dey, Lopamudra; Chakraborty, Sanjay; Biswas, Anuraag; Bose, Beep; Tiwari, Sweta," Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier", eprint arXiv:1610.09982,ARXIV, Volume-8, Issue-4, 2016,pp.54-62.
6. Md Shad Akhtar, Ayush Kumar, Asif Ekbal, Pushpak Bhattacharyya," A Hybrid Deep Learning Architecture for Sentiment Analysis"shad.pcs15,ayush.cs12,asif,pb}@iitp.ac.in.
7. Appel O., Chiclana F., Carter J., Fujita H."A Hybrid Approach to Sentiment Analysis with Benchmarking Results",Trends in Applied Knowledge-Based Systems and Data Science. IEA/AIE 2016. Lecture Notes in Computer Science, vol 9799. Springer, Cham.
8. Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto," SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods" EPJ Data Science (2016) pp.5-23.
9. Ahmed Abbasi, Ammar Hassan, Milan Dhar," Benchmarking Twitter Sentiment Analysis Tools,LREC,2014,pp-823-829.
10. Abhijit Chakankar, Sanjukta Pal, Mathur Krishna Venuturimilli," Sentiment analysis of users ' reviews and comment",<http://cs229.s tanford.edu/proj 2012>
11. Vivek Narayanan, Ishan Arora, Arjun Bhatia,"Fast and accurate sentiment classification using an enhanced Naive Bayes model. "Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2013 - Volume 8206,pp. 194-201.
12. K.Bar," SentimentAnalysisof Movie Reviews andTwitter Statuses", <http://www.cs.tau.ac.il/~kfirbar/mlproject/project.html>.
13. Bing Liu," Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkha and F. J. Damerau), 2010,pp1-18.
14. Peter D. Turney," Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424..
15. Prateek Garg," Sentiment Analysis of Twitter Data using NLTK in Python",M.tech,thesis,THAPAR UNIVERSITYPATIALA,2016.
16. S.K.Singh, S.Paul,and D.Kumar,"Sentiment Analysis,Approaches on Different Data Set domains"International Journal on Database Theory and Applications,Vol 7,No5,(2014)pp.39=50.
17. Md. Daiyan1, Dr. S. K. Tiwari2, Manish Kumar3, M. Aftab Alam," A Literature Review on Opinion Mining and Sentiment Analysis", IJETAE,Volume 5, Issue 4, April 2015,pp262-280.
18. Havinash P.H,Jeril Johnson N, Glen Thomas,Emily Stephen P," Mining Opinion Features in Customer Reviews",IJCERT, Volume 3, Issue 9, September-2016, pp. 535-539.
19. Xinmiao Li, Jing Li, Yukeng Wu, A Global Optimization Approach to Multi-Polarity Sentiment Analysis., PLoS ONE 10(4): 2015.pp1-18
20. Gurneet Kaur, Abhinash Singla," Sentimental Analysis of Flipkart reviews using Naive Bayes and Decision Tree algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016,pp148-153.
21. N. Roy and S.Some," Product Aspect Extraction for Sentiment Analysis without using Parsers" CS671 project report, IIT,Kolkata,India.
22. N.Chopra,P.kumar,R.Gupta, "Automated extraction of product attributes from reviews" CS 224N Final Project,2009.

23. S.Mukherjee and P.Bhattacharya," Feature Specific Sentiment Analysis for Product Reviews",Final project report IIT,Bombay, {subhabratam, pb1@cse.iitb.ac.in.
24. M.S.ElliandY.F.Wang,"AmazonReviews,businessanalyticswithsentimentanalysis"homes.soic.indiana.edu/mselli/SocialMedia Mining.pdf,
25. P.D.Turney,"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.
26. H.Cui,V.Mittal,and M.Datar," Comparative Experiments on Sentiment Classification for Online Product Reviews", Proceeding AAAI'06 proceedings of the 21st national conference on Artificial intelligence Volume 2,2006, pp. 1265-1270.
27. Richa.S,Swetha.N and Rekha.J," Mining of Product Reviews at Aspect Level", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014,pp87-95.
28. Akshay,M. and Navajyoti,S. "Generating domain specific sentiment lexicons using the Web Directory", Advanced Computing: An International Journal (ACIJ), Vol.3, No.5, September 2012,pp.45-51.
29. T. Wilson,J.Wiebe,and P.Hoffmann,"Recognizing Contextual Polarity:An Exploration of Features for Phrase-Level Sentiment Analysis", journal = {Computational Linguistics,2009,pp.399-433
30. G.Kaur and A.Singhla,"Sentimental Analysis of Flipkart reviews usingNaïve Bayes and Decision Tree algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016,pp.148-153.
31. S.Poria,E.Cambria,L.W.Ku,C.Gui,A.Gelbukh," A Rule-Based Approach to Aspect Extraction from Product Reviews",Unpublished report
32. F.J.M.Arboleda,G.A.A. Velasquez,G.L.P.Jimenez," A Proposal for Brand Analysis with Opinion Mining",Ch-4, E-Business - State of the Art of ICT Based Challenges and Solutions,pp63-80
33. Qian Liu1, Bing Liu, Yuanlin Zhang, Doo Soon Kim5 and Zhiqiang Gao," Improving Opinion Aspect Extraction Using Semantic Similarity and Aspect Associations", Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16),pp.2986-2992/
34. Y.Meng," Sentiment AnalysisA Studyon Product Features", Thesis Dessertation,4-2012,University of Nebraska.
35. Elliot Mars, Zachary,Y.F," Aspect Specific Sentiment Analysis of UnstructuredOnlineReviews",<https://cs224d.stanford.edu/reports/MarxElliot.pdf>
36. M.Hu and B.Liu," Mining Opinion Features in Customer Reviews", American Association for Artificial Intelligence,2004,
37. X.Fang and J.Zahn, "Sentiment analysis using product review data", Journal of Big Data (2015) 2:5,pp.1-14
38. A.Chakankar, S.P.Mathur, and K.Venturimilli, "Sentiment analysis of] users' reviews and comments"cs229, Stanford university, 2012
39. S.M.Babu and S.N.Das," An Unsupervised Approach for Feature Based Sentiment Analysis of Product Reviews", International Journal of Scientific Research Engineering & Technology (IJSRET),Vol.4,Issue.1, May, 2015. pp.484-489.

